Activity with Weka

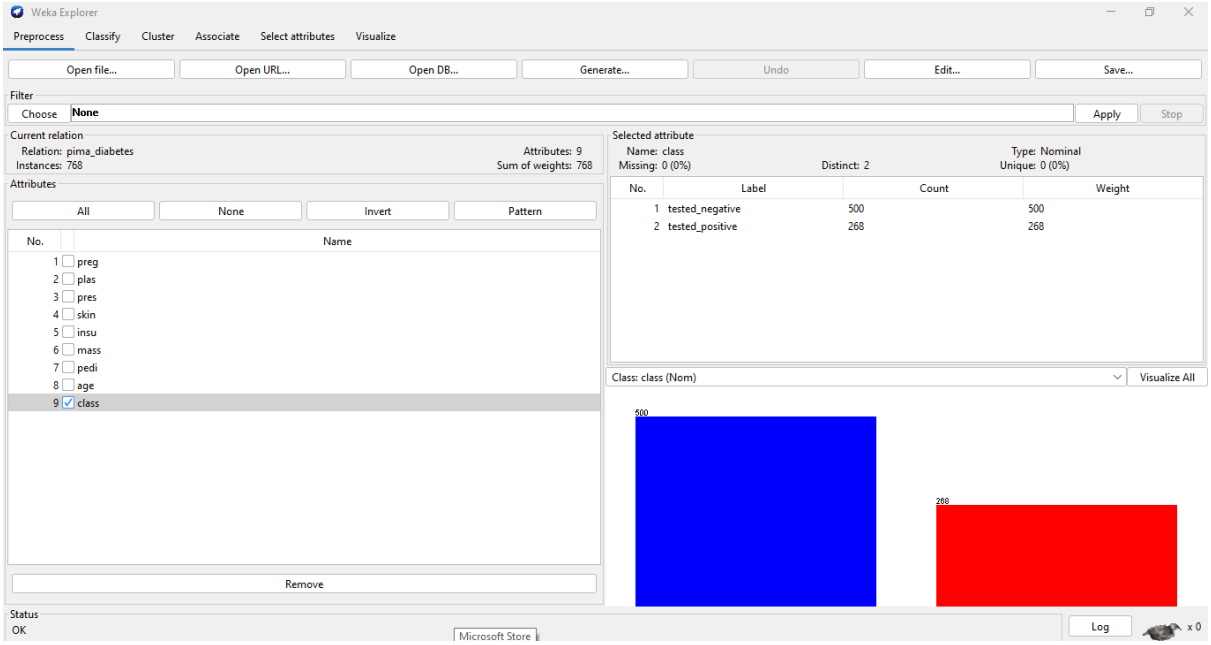
Week - 8

By Pulkit Kashyap

Assignment Activity Objective: KNN Model Evaluation

Practical Assignment

Task: Using the diabetes.arff dataset, apply KNN with different values of k (k=1, k=3, k=5). Compare the accuracy, precision, and recall for each value of k. Discuss the effect of changing k on the model's performance.



K = 1

```
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         539               70.1823 %
Incorrectly Classified Instances       229               29.8177 %
Kappa statistic                          0.3304
Mean absolute error                      0.2988
Root mean squared error                  0.5453
Relative absolute error                 65.7327 %
Root relative squared error            114.3977 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.794    0.470    0.759      0.794    0.776      0.331    0.650     0.732     tested_negative
                 0.530    0.206    0.580      0.530    0.554      0.331    0.650     0.469     tested_positive
Weighted Avg.    0.702    0.378    0.696      0.702    0.698      0.331    0.650     0.640

=== Confusion Matrix ===

   a   b   <-- classified as
 397 103 |   a = tested_negative
 126 142 |   b = tested_positive
```

K = 3

```
IB1 instance-based classifier
using 3 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         558               72.6563 %
Incorrectly Classified Instances       210               27.3438 %
Kappa statistic                          0.3822
Mean absolute error                      0.3092
Root mean squared error                  0.4525
Relative absolute error                 68.0324 %
Root relative squared error             94.9365 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.820    0.448    0.774      0.820    0.796      0.384    0.742     0.804     tested_negative
                 0.552    0.180    0.622      0.552    0.585      0.384    0.742     0.569     tested_positive
Weighted Avg.    0.727    0.354    0.721      0.727    0.722      0.384    0.742     0.722

=== Confusion Matrix ===

   a   b   <-- classified as
 410  90 |   a = tested_negative
 120 148 |   b = tested_positive
```

K = 5

```
IB1 instance-based classifier
using 5 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         562               73.1771 %
Incorrectly Classified Instances       206               26.8229 %
Kappa statistic                          0.3874
Mean absolute error                      0.3165
Root mean squared error                  0.4318
Relative absolute error                 69.6387 %
Root relative squared error             90.5982 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.836     0.463     0.771      0.836    0.802      0.390    0.766     0.828     tested_negative
               0.537     0.164     0.637      0.537    0.583      0.390    0.766     0.619     tested_positive
Weighted Avg.  0.732     0.358     0.724      0.732    0.726      0.390    0.766     0.755

=== Confusion Matrix ===

   a   b   <-- classified as
 418  82 |   a = tested_negative
 124 144 |   b = tested_positive
```

Comparison table

K Valvue	Accuracy (%)	Precision	Recall
K = 1	70.18	0.696	0.702
K = 3	72.65	0.721	0.727
K = 5	73.17	0.724	0.732

Analysis and Discussion

The data shows how the hyperparameter **k** affects the performance of the **K-Nearest Neighbors (KNN)** model.

1. Trend from k=1 to k=5

As **k** increases from 1 to 5, all performance metrics improve:

- Accuracy increases from 70.18% (k=1) → 73.17% (k=5) → higher at k=5
- Precision and Recall also rise steadily with larger k values.

This happens because a very small k (like 1) makes the model too sensitive to noise and outliers (overfitting). Increasing k smooths the decision boundary, leading to more stable and reliable predictions.

2. Effect of Changing k

- **k=1:** Low bias, high variance — fits the training data too closely and reacts strongly to noise.

- **k=3:** More balanced — better generalization and improved accuracy.

- **k=5:** High bias, low variance — the model is more generalized and less affected by random noise, giving the best performance here.

Conclusion

For this diabetes dataset, **k=5** provides the best overall results across accuracy, precision, and recall. indicating that the model generalizes better with a moderate neighborhood size

It offers a strong balance between fitting the data well and making it the most suitable choice among k=1, 3, and 5.